ECP PowerSteering Project

PI: Tapasya Patki* (Lawrence Livermore National Laboratory)

Barry Rountree (Co-PI, LLNL), Aniruddha Marathe (LLNL), David Lowenthal (University of Arizona), Samuel Cotter (University of Arizona), Scott Walker (University of Arizona), Jonathan Eastep (Intel)

November 14, 2018

LLNL-PRES-753368

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

EXASCALE COMPUTING PROJECT

National Nuclear Security Administration

U.S. DEPARTMENT OF ENERGY | Office of Science

Previously: Martin Schulz, FY17

Overview: Production-grade, open source, scalable runtime integrates into HPC PowerStack

Problem:

- Power and energy are critical constraints for exascale
- Inefficient power management results in limited application performance, job throughput and system utilization, leading to added operational costs
- Existing approaches are ad-hoc research codes (Conductor, Adagio, RMAP, etc.) and have several scalability and portability limitations

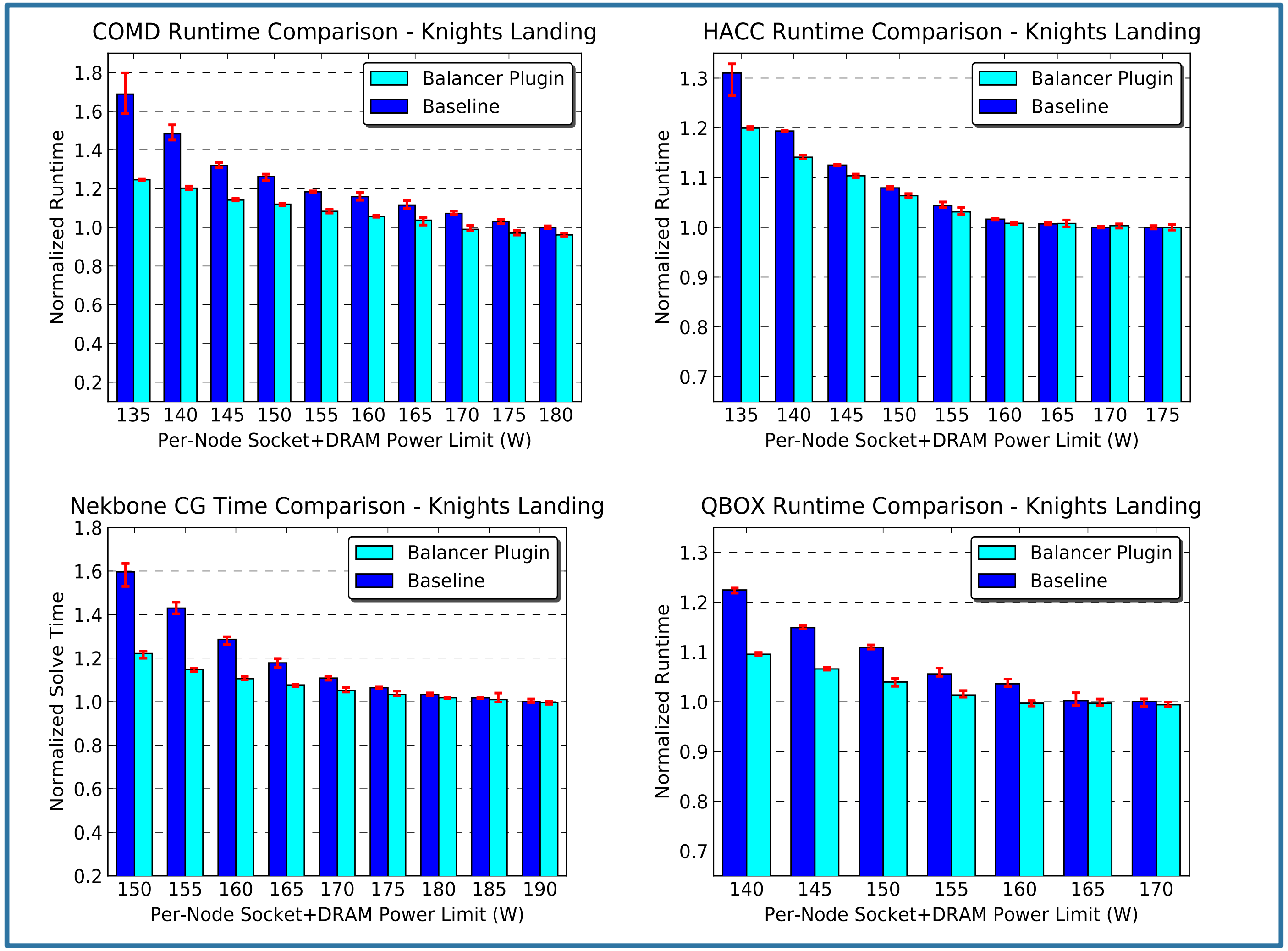
Solution:

- Production-grade, industry-supported, open-source, job-level runtime (GEOPM) suitable for integration with resource manager/software stack
- Algorithms to analyze critical path of applications, distribute power intelligently to hardware components, mitigate variation, support portability to upcoming architectures and task-based programming models

Impact goals and impact metrics

Impact Goal*	Metric
Widespread use of GEOPM across ECP-enabled applications.	Number of ECP benchmarks, scientific applications, system software components, and processor architectures that have been integrated with GEOPM.
Demonstrate safe execution under either power or energy constraints.	Using multiple benchmarks, proxy applications and applications, sample instantaneous power and measure total energy, and demonstrate system-specified bounds are not exceeded.
Optimize runtime in power and energy constrained environments, with an expected average improvement of 20%*.	Show percentage runtime performance improvement across a selected suite of multiple benchmarks, proxy applications and codes while maintaining power at or under the system-specified bound. Comparison will be made with naïve uniform static power allocation and/or with full-energy execution. (*Exact improvements will depend on underlying processor architecture and application characteristics).

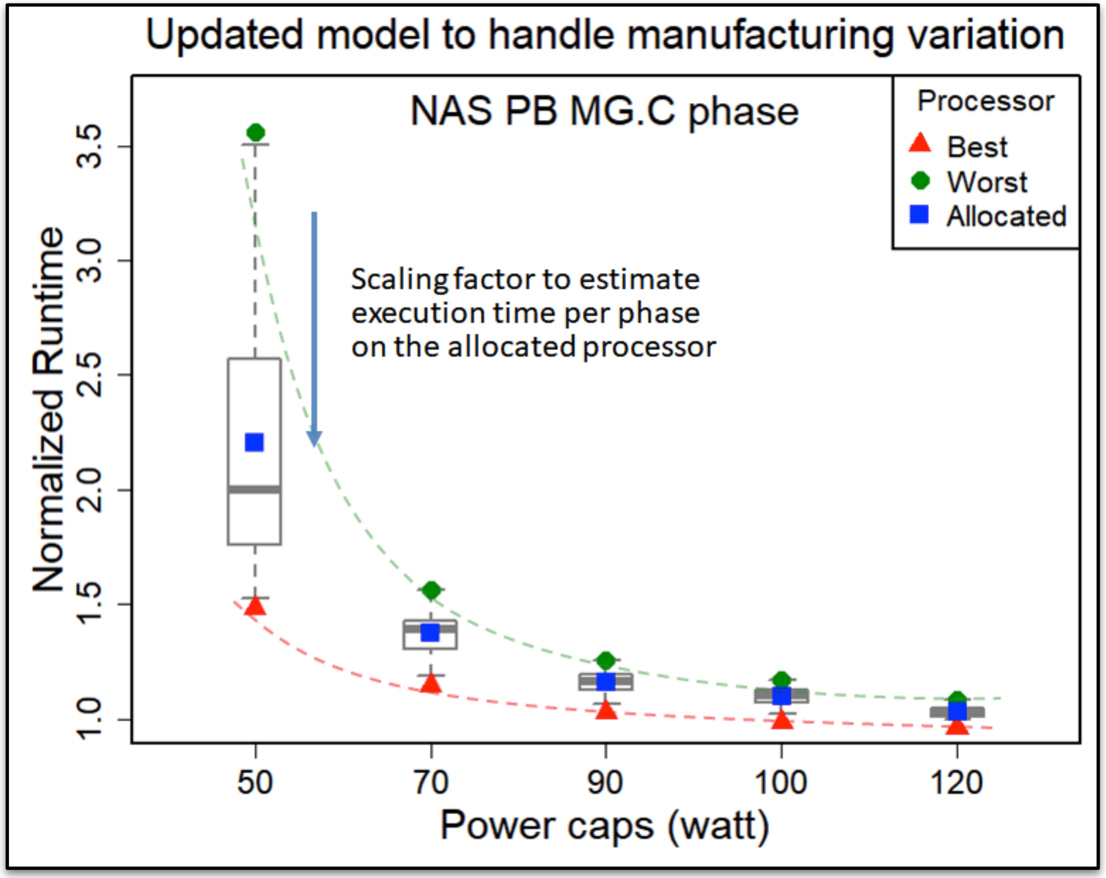
Power Steering can accomplish more science per dollar



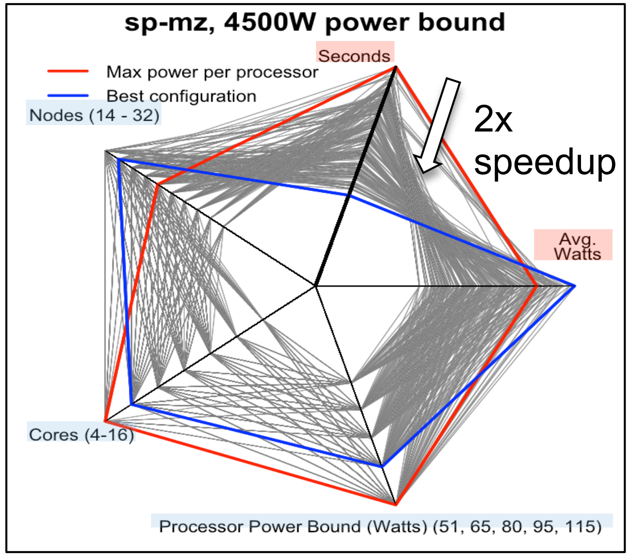
- Power Steering can improve Time To Solution (TTS) by up to 30% on ECP applications
- 30% improvement in TTS translates to 30% of power savings

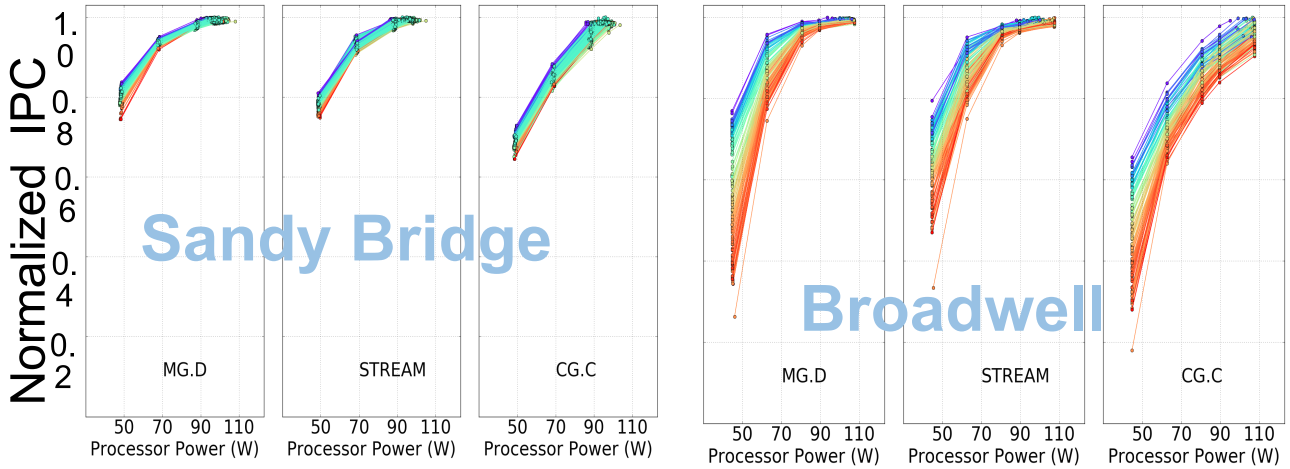
- Example:
  - If we assume a 30 MW system that is operational for 5 years, this is equivalent to 30 MW * 30% * 5, or *45 MW-Years*
  - Assuming a power cost of $1M per MW-year, that is *$45M for additional science*

New power model with configuration space exploration



- Select application configurations intelligently at runtime
- Address manufacturing variation with a non-linear model





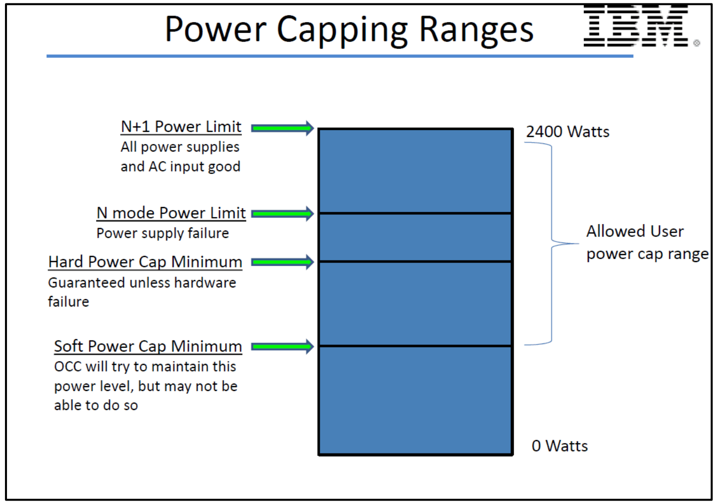
https://github.com/amarathe84/geopm/tree/dev/ecp

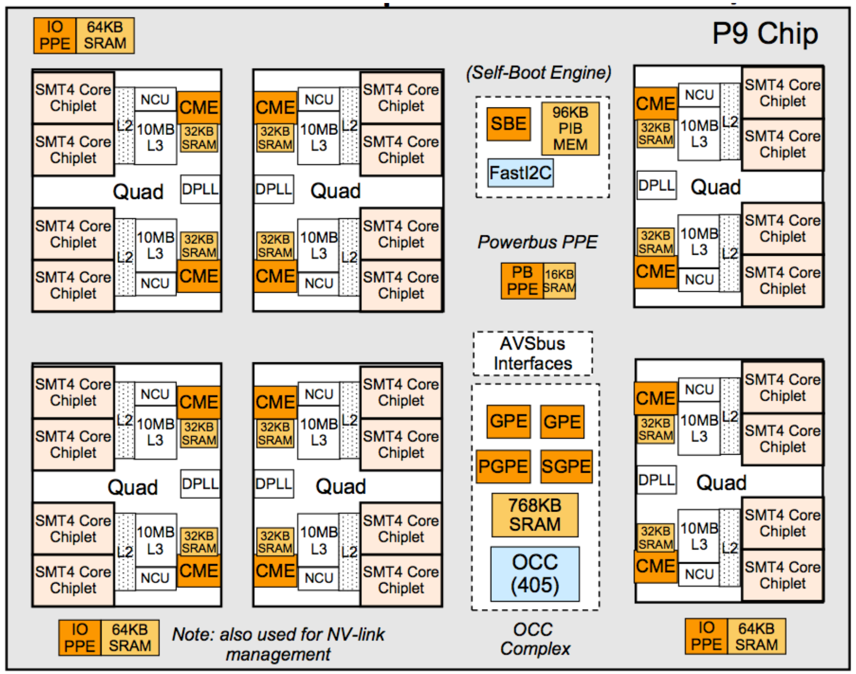
Port GEOPM to non-Intel architecture (IBM Power9)

- Purchased an IBM Power9 Witherspoon node for the Power Lab at LLNL
  - Allows for isolated root access, low level firmware development, disabling of features such as secure boot
  - Replica of a Sierra node, which allows developed software to be easily transferrable

- Developed DVFS-based model for GEOPM, explored OCC (on-chip controller) options

- Identified a bug in IBM OPAL firmware
  - Did not account for scenarios where GPUs were not used
  - Did not allow for setting of correct power caps
  - Did not expose knobs for TurboBoost/UltraScale

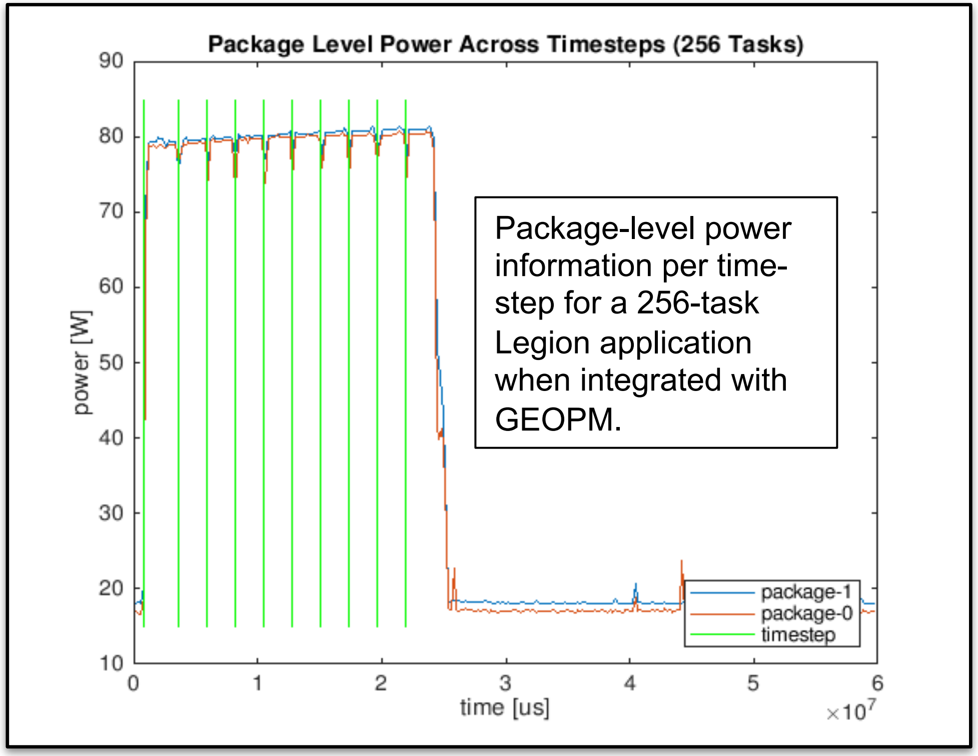


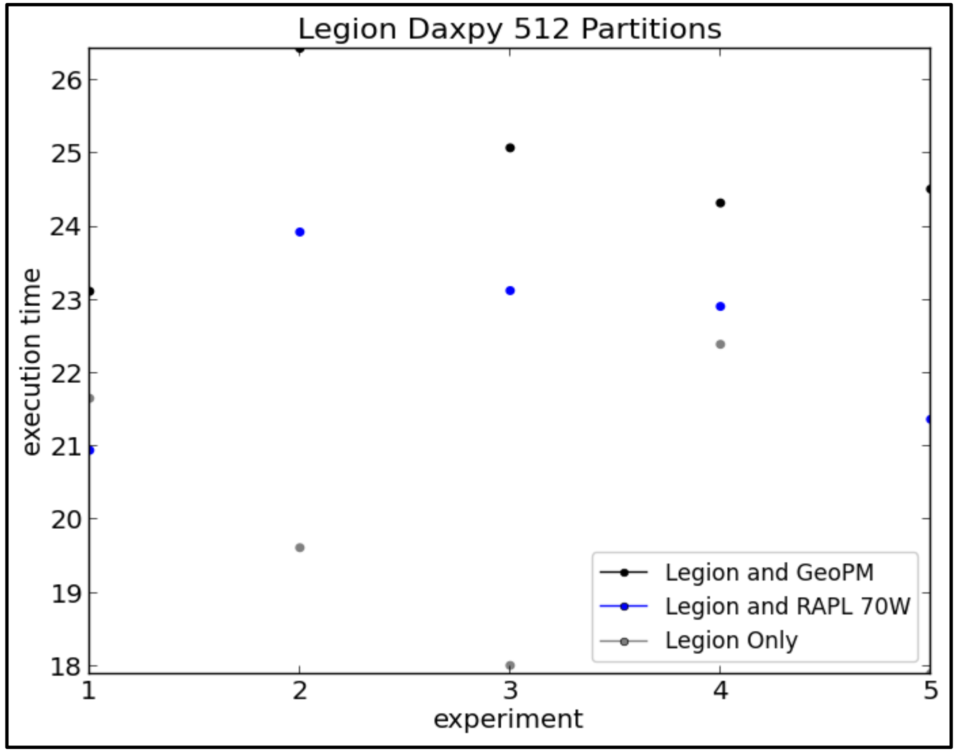


https://github.com/amarathe84/geopm/tree/ibm-port

https://github.com/open-power/skiboot/issues/195

Evaluate Legion applications, design power management for task-based models





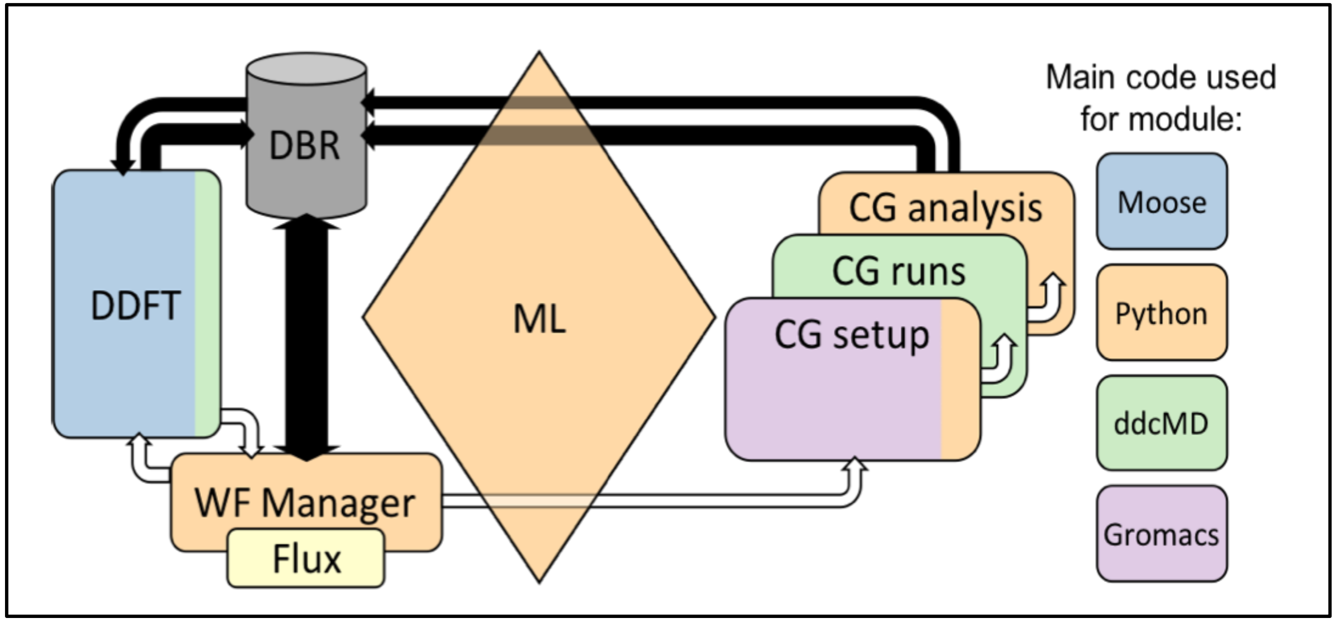
Experiments with the Legion DAXPY benchmark running without a power cap, with a 170W power cap with GEOPM, and with a 140W cap with RAPL. Execution time is shown on y-axis for 5 experiments.

- Successful integration of Legion and GEOPM, not implemented as a plugin due to MPI-related restrictions in current version of GEOPM
- Created a new DAXPY benchmark for evaluation

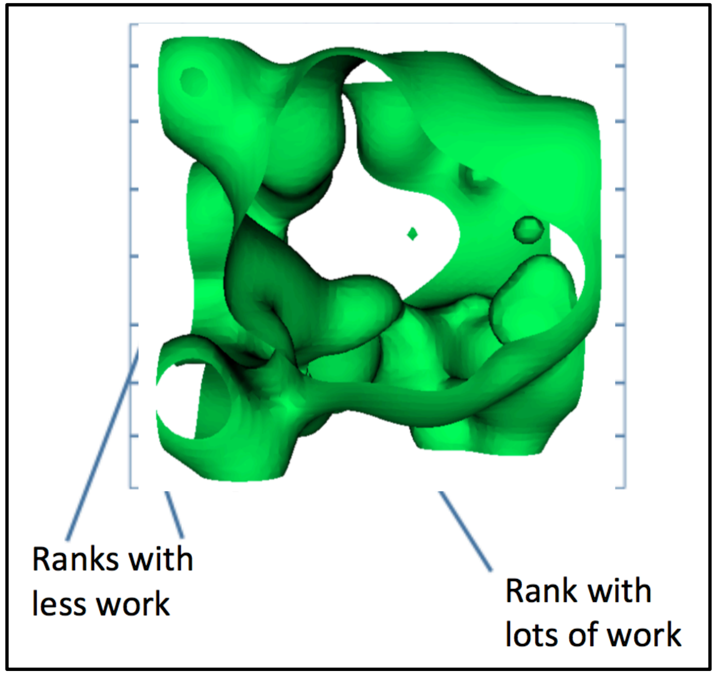
https://github.com/scott-walker-llnl/legion-geo-interop

Scientific workflows need fine-grained power management

RAS Cancer Simulation Workflow



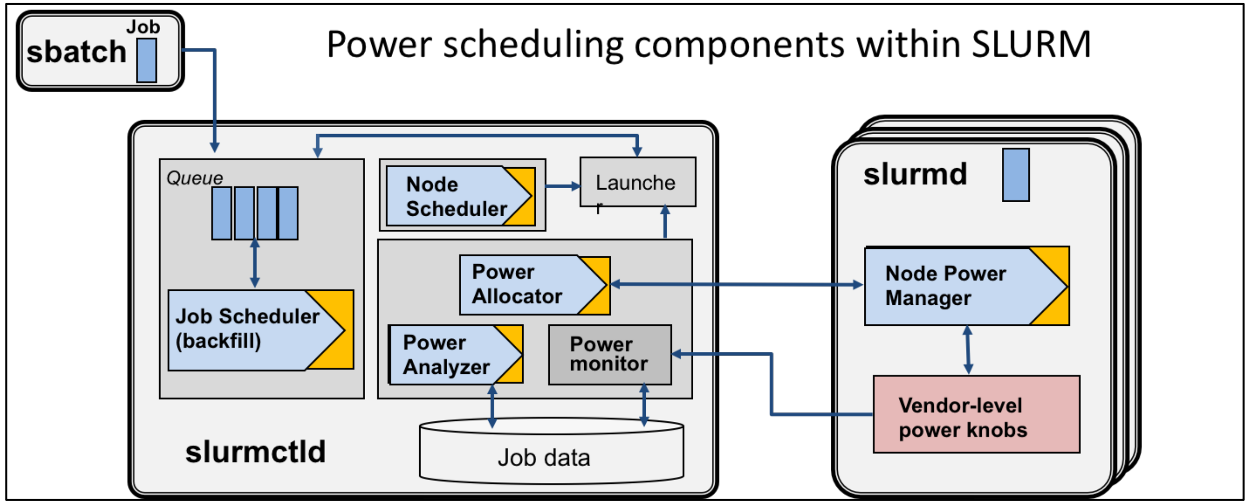
Isosurfacing and Visualizations

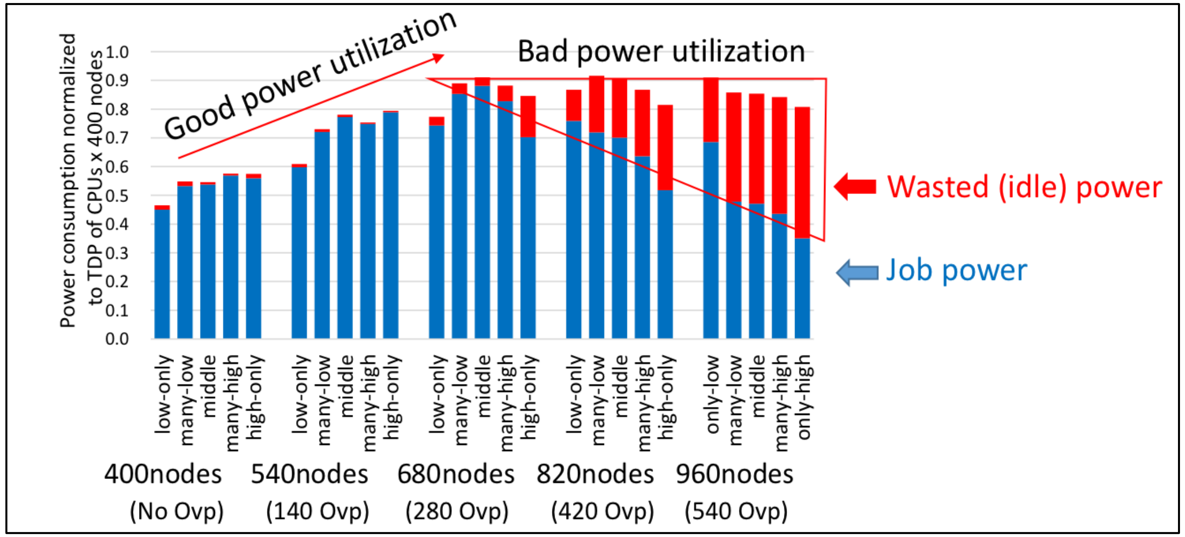


- Load imbalance cannot be addressed directly as memory may be shared between simulation, analysis and visualization components making data movement challenging
- Parts of large-scale workflows may not utilize GPUs or certain cores
- Critical path can be sped up by directing power to relevant tasks

EXASCALE COMPUTING PROJECT

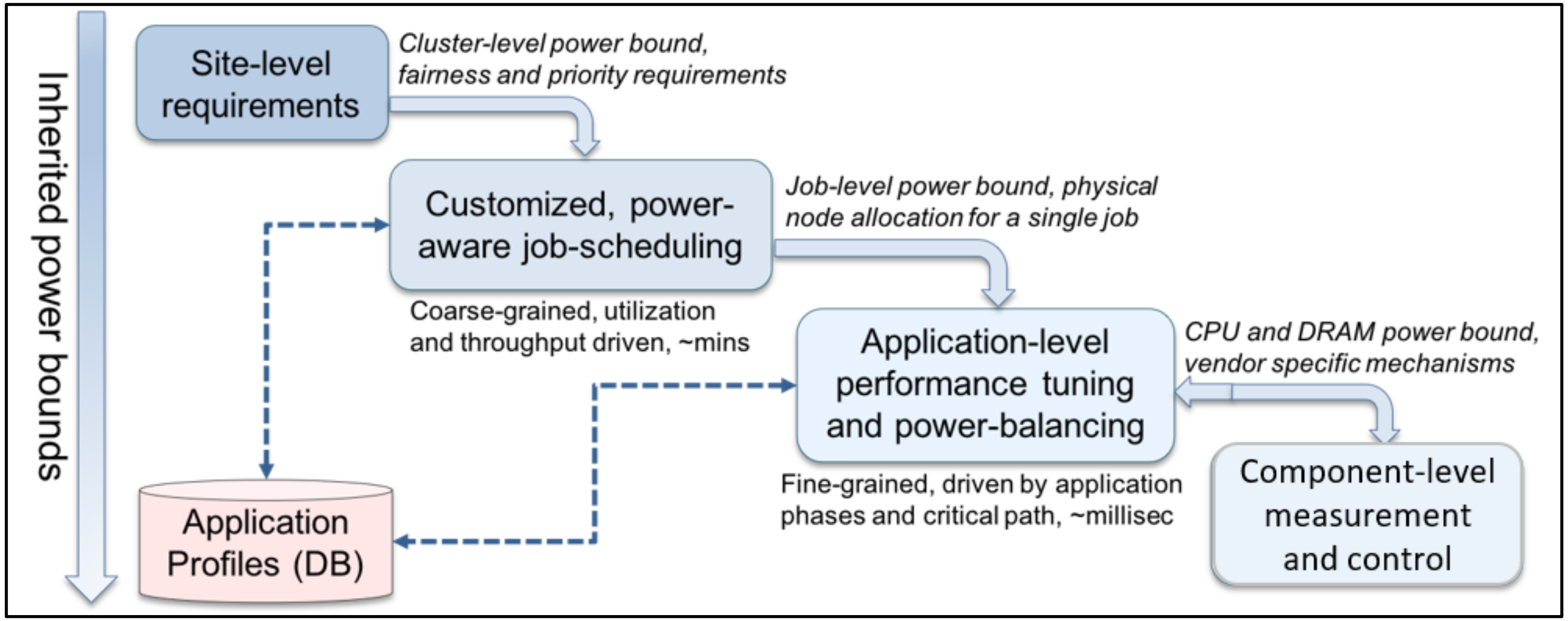
Explore interfaces for GEOPM and HPC batch schedulers for ECP Argo





- Implement and test power-aware SLURM at scale
- Explore interfaces for fine-grained management and identify range of improvement
- Five job mixes, 5 levels of overprovisioning to understand the impact of degree of overprovisioning
- IvyBridge cluster HA8K in Japan, 965 nodes
- Sweet spot around 680 nodes shows that hardware overprovisioning with GRM can give better utilization and up to 40% higher throughput

Summary and Next Steps



HPC PowerStack

https://powerstack.lrr.in.tum.de/

- We are collaborating with scientific workflow teams, engaging users and evaluating more ECP applications
- We are supporting multiple architectures and helping with community outreach through the HPC PowerStack charter

Thanks!